COVID-19 ANALYSIS USING THE GOMPERTZ FUNCTION

Sergi Pradas Rodriguez,[*] David Rovira Ferrer,[†] and Xavier Simó Romaguera[‡]
Projectes d'Enginyeria Física II.
Enginyeria Física.

(Dated: June 17, 2020)

This report analyses the spread of COVID-19 in european countries, focusing especially on Italy and Spain, for which short-term predictions are made for the cumulative number of hospitalizations, ICUs, discharges and deaths. Taking into account the different magnitudes considered, for 1 day predictions the mean probability of a right guess is of 0.99, 0.94 for 2 day predictions, 0.89 for 3 day predictions and 0.83 for 4 day predictions.

I. INTRODUCTION

COVID-19 is an infectious disease caused by a virus called *SARS-CoV-2*. It usually causes tiredness, fever and breathing problems. It is not often serious, but it can lead to severe problems, and ultimately death, for some people.

The objective of this project is to analyze, via the use of the Gompertz function, the evolution of the number of cases, hospitalizations, ICUs, discharges and deaths due to COVID-19, as well as make mid/long-term predictions based on current data. This analysis has been developed with the help of and close collaboration with the BIOCOM-SC group from UPC, which has been working on the study and evolution of COVID-19 since the first days of the epidemic. These results, among many more regarding the spread of the disease, have been sent daily to the European Commission for three months.

This report is structured in the following way: in Section II A the Gompertz function and its characteristics are introduced. In Section II B the process of systematization of the information to be used in the predictions is explained. This is followed by an account of the data analysis of the significant magnitudes in Section II C. Then, in Section III A a study of the predicted final incidence for different countries is carried out, and in Section III B the accuracy and reliability of the predictions is analysed. Finally, in Section IV the conclusions are drawn.

II. METHODS

A. GOMPERTZ FUNCTION

The main objective of this project is to study the use of the Gompertz function to predict not only the future number of cases of COVID-19, as it has already been shown capable of doing [1], but to predict the spread of the disease in terms of hospitalized people, intensive care treatment units occupied (referred as ICUs henceforth), discharges and deaths.

This empirical model is characterized by an initial exponential growth that, overtime, slows down. It is, thus, able to portray the evolution and, later, control of the disease. It is described by the following equation:

$$N(t) = Ke^{-\ln\left(\frac{K}{N_0}\right)e^{-a(t-t_0)}} \tag{1}$$

Where *N(t)* is the cumulated number of cases at time t (measured in days), or the magnitude studied, N_0 is the value of such magnitude at time t_0, a marks the slowing down of the spreading's rate, and K is the expected final value of the variable under analysis at the end of the crisis. This model is fitted to the country or region being analyzed in order to find K and a. A detailed explanation of this process will be carried later on in this report.

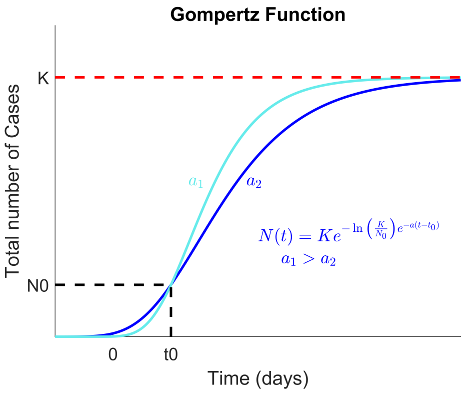


FIG. 1. This figure shows the Gompertz function. As it can be seen, the higher the a parameter is, the faster the spreading is.

It is, then, of major interest, to be able to predict when the maximum number of new cases in a day will be registered. This is done by analyzing the first derivative of the Gompertz function, as it represents the new cases each day:

$$\frac{dN}{dt} = aKe^{-\ln\left(\frac{K}{N_0}\right)e^{-a(t-t_0)}}\left(\ln\left(\frac{K}{N_0}\right)e^{-a(t-t_0)}\right) \tag{2}$$

Therefore, by computing the second derivative and setting it to zero we get the day of maximum number of new

[*] sergi.pradas@gmail.com
[†] drovfer@gmail.com
[‡] xavi.simo.99@gmail.com

cases, marked as t_{peak}:

$$t_{peak} = t_0 + \frac{\ln\left(\ln\left(\frac{K}{N_0}\right)\right)}{a} \tag{3}$$

We can then analyze the error in such prediction by propagating the error:

$$\sigma^2_{t_{peak}} = \left(\frac{\delta t_{peak}}{\delta a}\right)^2 \sigma_a^2 + \left(\frac{\delta t_{peak}}{\delta K}\right)^2 \sigma_K^2 \tag{4}$$

Where σ_a and σ_K are the error margins for a and K, respectively, as obtained by the fitting. Which, substituting, yields:

$$\sigma^2_{t_{peak}} = \left(\frac{-\ln\left(\ln\left(\frac{K}{N_0}\right)\right)}{a^2}\right)^2 \sigma_a^2 + \left(\frac{1}{aK\ln\left(\frac{K}{N_0}\right)}\right)^2 \sigma_K^2 \tag{5}$$

Moreover, it is of crucial interest to be able to predict when the spreading of the disease reaches a critical point regarding the final expected number of cases. We take that critical point to be 90% of K, as it is then when governments can start to consider relaxing their measures of control. In order to compute when that happens, we need to consider $N(t_{90}) = 0.9K$, with t_{90} the day when the 90% of the expected final number of cases is reached. Substituting into the Gompertz function yields:

$$t_{90} = t_0 - \frac{\ln\left(\frac{-\ln(0.9)}{\ln\left(\frac{K}{N_0}\right)}\right)}{a} \tag{6}$$

The error in the prediction of this value is:

$$\sigma^2_{90} = \left(\frac{\partial t_{90}}{\partial K}\right)^2 \sigma_K^2 + \left(\frac{\partial t_{90}}{\partial a}\right)^2 \sigma_a^2 \tag{7}$$

With:

$$\frac{\partial t_{90}}{\partial K} = -\frac{1}{aK\ln\left(K/N_0\right)} \;\&\; \frac{\partial t_{90}}{\partial a} = -\frac{\ln\left(\frac{-\ln(0.9)}{\ln(K/N_0)}\right)}{a^2} \tag{8}$$

B. RETRIEVAL OF DATA AND SYSTEMATIZATION

Given the main objective of this project is to study the extension of the Gompertz's function fitting to other variables apart from the number of cases, such as the number of hospitalizations and ICUs, among others, we had to work on finding the information required for the analysis itself. In this sense we worked for several days on finding reliable and useful information about the magnitudes mentioned above for different countries in the European Union. By useful information it is meant data in the form of *csv* or *excel* files that would make an systematization of the actualization process viable. This proved to be more difficult than first anticipated, as the majority of official ministerial webpages showed the data in the form of graphical analysis as public release and were not meant for use in further studies.

In the end, it was possible to find such reliable information, with varying success, for three EU countries: Belgium[2] (including its regions), France[3], Sweden[4] and Switzerland[5]. The analysis of Sweden is of major interest, as it can show the effects of different control measures, as Sweden has chosen to follow less restrictive actions.

Moreover, we were given the necessary data regarding Italy[6] and Spain[7] (and their respective regions and communities) by the BIOCOM-SC group from UPC. To this, one has to add the data regarding cases and deaths collected by the European Center for Disease Prevention and Control[8] (ECDC), that will make possible a supranational analysis for European countries.

However, given that every country organizes its data in different forms, be it by regions, days, sex or even age, an homogenization of the information was required. This meant that a *Matlab* code had to be developed, that would take into account those differences, and would generate a file with the same format as those provided by the BIOCOM-SC group. For Belgium this proved to be a complete success, as a retrieval of information regarding cases, hospitalizations, ICUs, discharges and deaths proved possible. The same can be said of France, except obtaining information of the separate regions proved more difficult and could not be done, as opposed to Belgium. In the case of Sweden we were not able to find data in useful form regarding hospitalizations and discharges. On the other hand, some problems started to appear as soon as we started trying to make predictions:

1. Belgium: ICUs are prevalence values. This means that we only have access to the current occupation, so we lack information about the cumulative values.
2. France: Hospitalizations and ICUs are prevalence values, and not cumulative.
3. Italy: Hospitalizations and ICUs are prevalence values.
4. Spain: More general problems appeared due to inconsistencies and mistakes in the reported data.

This meant that for the data presented in prevalence form no prediction was possible, as predictions are based on cumulative values.

C. DATA ANALYSIS OF SIGNIFICANT MAGNITUDES

In this section we present the analysis of cases, hospitalizations, ICUs, discharges and deaths due to COVID-19 carried out for different countries and regions within them. The main objective is to make short-term predictions that are useful for decision makers within governments and other state institutions.

The model is adjusted to make predictions for the cumulated values of the magnitudes indicated above when certain conditions are met regarding the reported information:

1. Cases: At least 3 days with more than 100 cases and 1 day with more than 200.
2. Hospitalizations: At least 3 days with more than 50 hospitalizations and 1 day with more than 70.
3. ICUs: At least 3 days with more than 10 ICUs and 1 day with more than 15.
4. Discharges: At least 3 days with more than 10 discharges and 1 day with more than 15.
5. Deaths: At least 3 days with more than 10 deaths and 1 day with more than 15 deaths.

In case any of these criteria is not met, no predictions are made for that variable. This fact is necessary to differentiate the phase in which the spreading of the disease is due to imported cases and the subsequent period in which new cases occur because of local transmission.

The prediction process is based on a Non-Linear Least Squares fitting of the parameters K and a of the Gompertz function using the Levenberg-Marquardt and the Trust-region algorithms. For this fitting only the days for which the criteria above are met are used in the process. In case there are more than 15 points that verify it, only the last 15 are used. In order to better capture the current trend of the spreading of the disease, if there are more than 9 points that meet the conditions, weights are applied to the last 3 points in the fitting process. Then, if there are more than 6 points verifying the criteria, predictions are made for the following 2 days; if there are more than 9, for the following 3 days; if there are more than 12, for the following 4 days and, if there are more than 15, for the following 5 days.

Moreover, we plot a variable we name ρ, that is related to the reproduction number, that is, the number of new infections caused by a single case. It is evaluated as follows for the day before the last reported information:

$$\rho(t-1) = \frac{N_{new}(t) + N_{new}(t-1) + N_{new}(t-2)}{N_{new}(t-5) + N_{new}(t-6) + N_{new}(t-7)} \tag{9}$$

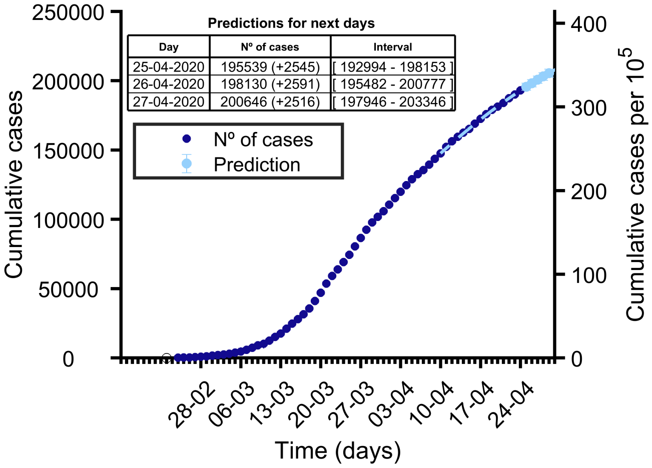
Where $N_{new}(t)$ is the number of new confirmed cases at day t.

As an example of the results obtained, in Figure (2) and (3) these are shown for Italy as a whole. The complete reports for every country are shown in a shared *Dropbox* folder [9].

III. ANALYSIS AND RESULTS

A. PREDICTION OF INCIDENCE AND RELATED MAGNITUDES

In this section we show the results from the predictions made for the day of maximum number of new cases, the day t_{90}, as defined above, and compare the current situation to the final expected situation by using the incidence,



Predictions for next days

Day	Nº of cases	Interval
25-04-2020	195539 (+2545)	[192994 - 198153]
26-04-2020	198130 (+2591)	[195482 - 200777]
27-04-2020	200646 (+2516)	[197946 - 203346]

FIG. 2. Total number of cases in Italy on April 24th. In navy blue the reported number of cases, and in light blue the predictions made. As it can be seen, the spread of the disease was not exponential anymore, and was starting to slow down.

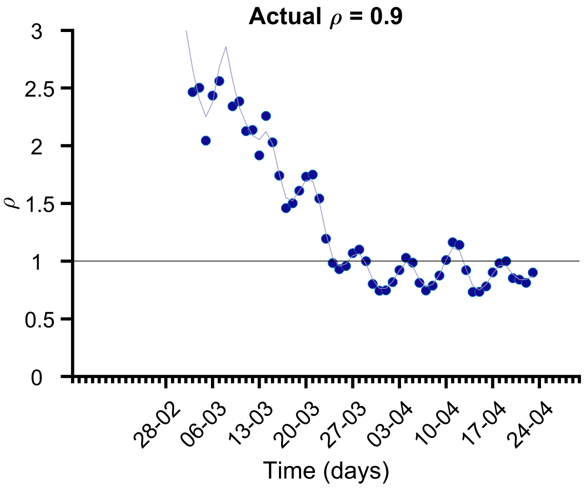


FIG. 3. $\rho's$ evolution in Italy as of April 24th. A black line marks $\rho = 1$, that indicates when the spread slows down.

that is, the number of cases for every 100000 citizens. This has been done for Italian and Spanish regions [9], as well as for EU and EFTA (European Free Trade Association) countries, and the UK. This last analysis is shown in Figure (4).

Moreover, a study of the convergence of t_{peak}, t_{90} and K, is done for Spain, Italy, France, Germany and the UK [9]. As an example, we show the results for Germany in Figure (5).

An analysis of the convergence of K is also carried out for Italian regions and Spanish autonomous communities [9]. For the case of EU-EFTA-UK as a whole the results are shown in Figure (6).

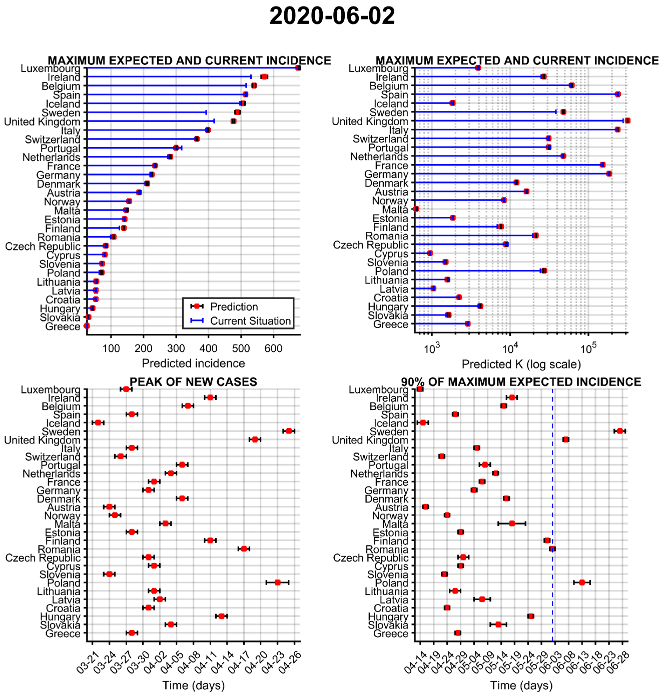


FIG. 4. Comparison of the current and final predicted situation for EU-EFTA-UK countries as of June 2nd.

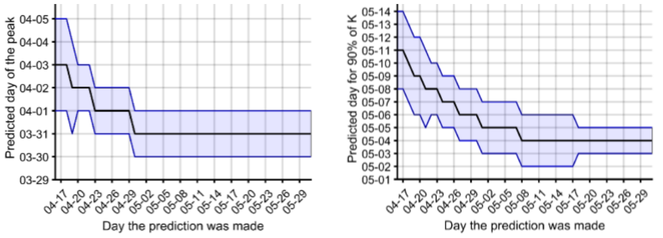


FIG. 5. Convergence of t_{peak} and t_{90} for Germany as of May 30th.

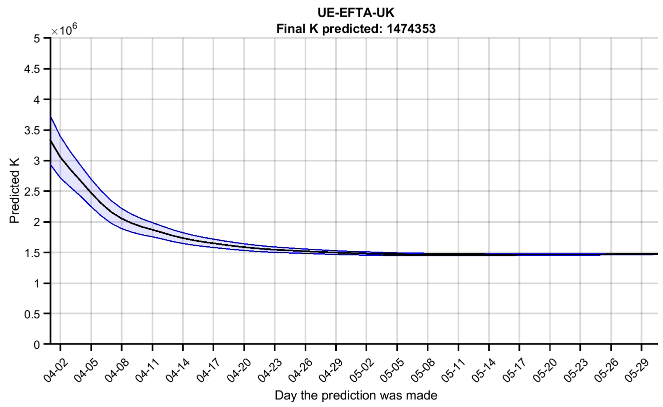


FIG. 6. Convergence of K for EU-EFTA-UK as a whole as of May 30th.

B. ACCURACY OF THE PREDICTIONS

In this section we present the analysis of the accuracy of the predictions shown in the previous sections for Belgium, France, Italy, Spain and Sweden, including not only the countries, but also the predictions for Belgian, Italian and Spanish regions or autonomous communities. For the purpose of these calculations only the variables in each country and region that are reported as cumulative values are used.

In order to compute this we store the predictions made each day for every variable, as well as the corresponding 99% confidence intervals. Then, that prediction is compared to the real values as reported by official sources. In case the latter values fall into the predicted confidence intervals we consider the prediction to be right, in the case it does not, we consider it to be wrong; this will be used to determine the reliability of the predictions. The exact results are shown in Table I.

PROBABILITY				
Variables	1 day prediction	2 days	3 days	4 days
Hospitalizations	0.98	0.94	0.88	0.83
ICUs	0.99	0.95	0.91	0.88
Discharges	0.99	0.92	0.85	0.78
Deaths	0.99	0.96	0.90	0.84

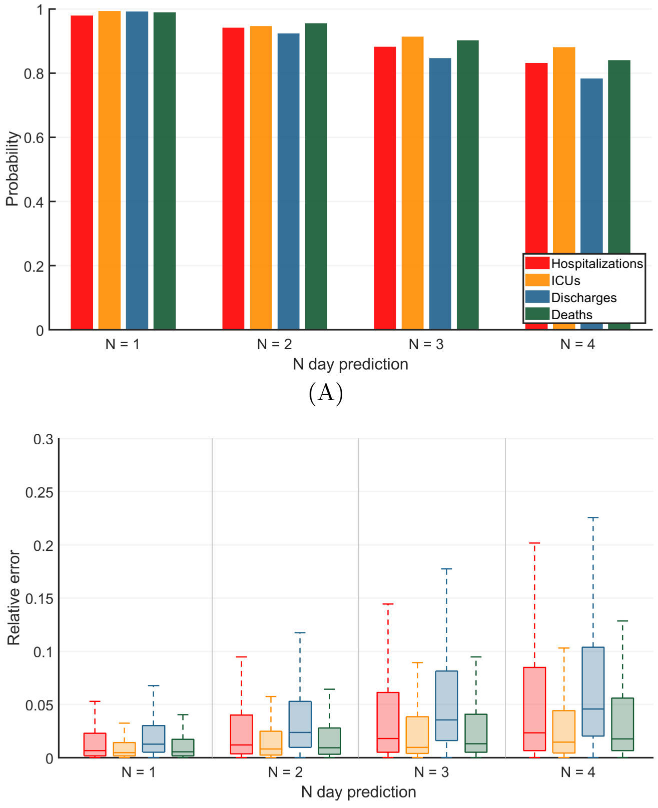
TABLE I. Reliability of the predictions made.

Moreover, we compute the relative error as (reported value - predicted)/(reported value). In order to show these results we plot them via a boxplot for the relative error, and a bar graph for the probability of making a correct prediction. The findings regarding hospitalizations, ICUs, discharges and deaths are shown in Figure (7).

For both the probability and the relative error the number of samples used is shown in Table II. For hospitalizations data from Belgium, Spain's autonomous communities and Switzerland is used due to other countries not providing the corresponding cumulative numbers. With respect to ICUs, data from Spain and Sweden is used. Regarding discharges data from Belgium and its regions, Italy and its regions and France is used. Information from Spain is not used due to problems with the historical series. For deaths, data from Belgium, France, Italy, Spain, Sweden and Switzerland is considered.

SAMPLES USED				
Variables	1 day prediction	2 days	3 days	4 days
Hospitalizations	1172	1152	1132	1112
ICUs	261	257	253	249
Discharges	756	744	732	720
Deaths	1619	1594	1569	1544

TABLE II. Samples used in the predictions.



(A)

(B)

FIG. 7. Errors of the predictions for hospitalizations, ICUs, discharges and deaths for up to 4-day predictions. $N = 1$ refers to the prediction made for one day after the day it was made. The same goes for $N = 2$, 3 and 4. (A) Probability that the predictions made fall in the prediction interval for each day. (B) Relative errors of the predictions.

We conclude that the relative errors for 1 and 2 day predictions are well below 5%, indicating the reliability of the predictions. For 3 and 4 day predictions ICUs yield more than 5% relative error, and for 4 day predictions deaths also reach this threshold. As can be seen from the previous figures, discharges are an outlier, as they show significantly higher relative errors, even reaching 10% for 4 day predictions. As will be explained in the following section we suspect this is due to the ”weekend effect”.

IV. CONCLUSIONS

The main objective of this report was to analyze the spread of COVID-19 via the use of the Gompertz function, and to make short-term predictions based on its fitting to the data provided by the different institutions involved in the monitoring of the disease. As the results provided in the last section show, the mean relative error for these short-term predictions is well below 10%, thus showing the potential of this function as a guide for policy makers. However, this approach suffers from some problems.

First of all, as is the general case for every prediction model, it relies entirely on the quality of the data that is made public, thus suffering from inconsistencies and mistakes in the daily publication of the data. One such hurdle is the effect known as ”weekend effect”, where lower number of cases, hospitalizations and other magnitudes are reported. This significantly modifies the day-to-day behaviour and influences the predictions that are made based on the fitting to the reported values. This has proven to be specially significant in the case of reported discharges, where major differences appear between workdays and weekends.

Moreover, other problems arise due to inconsistencies in the data reported by some governments and institutions. This is the case, for example, of Italy and Spain. Concerning the former, problems regarding discharges are specially important. For the latter, multiple inconsistencies have appeared during the last months. These include changes in the criteria employed in the reporting of the data of different magnitudes, such as changes in what is considered a case of COVID-19 and the inability of correcting the historical series for some autonomous communities. An example of these problems is shown in Figure (8).

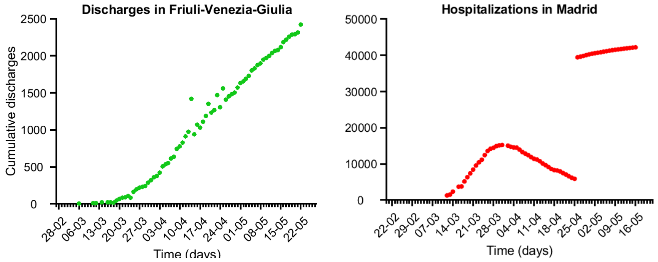


FIG. 8. (A) Cumulative discharges in Friuli-Venezia-Giulia, Italy. (B) Hospitalizations in Madrid, Spain. Both show clear problems in the historical series.

To conclude, the Gompertz function has proven to be a very useful tool in analyzing the short-term the behaviour of the spread of COVID-19, provided the quality of the data and the historical series allow for a good fitting of the function.

V. FINAL COMMENTS

We would like to thank the BIOCOM-SC research group from UPC, and especially Martí Català and Clara Prats, for the constant help with the day-to-day problems regarding the calculations and analyses.

REFERENCES

[1] CATALA, Marti, et al. Empiric model for short-time prediction of COVID-19 spreading. *medRxiv*, 2020.

[2] Sciensano, the Belgian institute for health. (2020). Epidemiological follow-up of the COVID-19 disease in Belgium. Retrieved from *https://epistat.wiv-isp.be/covid/*

[3] French Department of Public Health. (2020). Hospitalizations data regarding COVID-19 epidemic. Retrieved from *https://www.data.gouv.fr/fr/datasets/donnees-hospitalieres-relatives-a-lepidemie-de-covid-19/#_*

[4] Public Health Agency of Sweden. (2020). Historic data on the outbreak of coronavirus disease. Retrieved from *https://www.folkhalsomyndigheten.se/smittskydd-beredskap/utbrott/aktuella-utbrott/covid-19/bekraftade-fall-i-sverige/*

[5] Swiss Federal Office of Public Health. (2020). Data on confirmed cases of coronavirus infections and deaths due to the disease. Retrieved from *https://www.bag.admin.ch/bag/en/home/krankheiten/ausbrueche-epidemien-pandemien/aktuelle-ausbrueche-epidemien/novel-cov/situation-schweiz-und-international.html#1905396240*

[6] Italian Department of Civil Protection. (2020). Historic data on the outbreak of coronavirus disease in Italy. Retrieved from *https://github.com/pcm-dpc/COVID-19*

[7] Spanish Health Department and Carlos III Health Institute. (2020). Historic data on the outbreak of coronavirus disease in Spain. Retrieved from *https://github.com/datadista/datasets/tree/master/COVID%2019*

[8] European Centre for Disease Prevention and Control. (2020). Number of new cases reported per day and per country. Retrieved from *https://www.ecdc.europa.eu/en/publications-data/download-todays-data-geographic-distribution-covid-19-cases-worldwide*

[9] *https://www.dropbox.com/sh/vfok5ii8gawsrkp/AACi3xCxp-ibXFyvIzzlipwCa?dl=0*